

Update on Dataverse



Image credit: David Bygott (CC-BY-NC-SA)

2014 Dryad-Dataverse Community Meeting

Mercè Crosas, Elizabeth Quigley & Eleni Castro

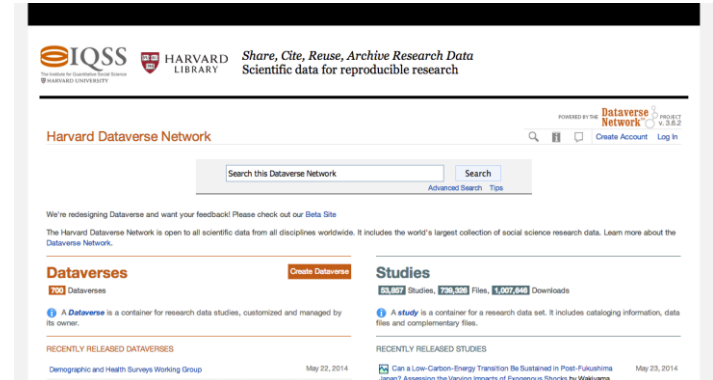
Data Science > IQSS > Harvard University

Introduction to Dataverse

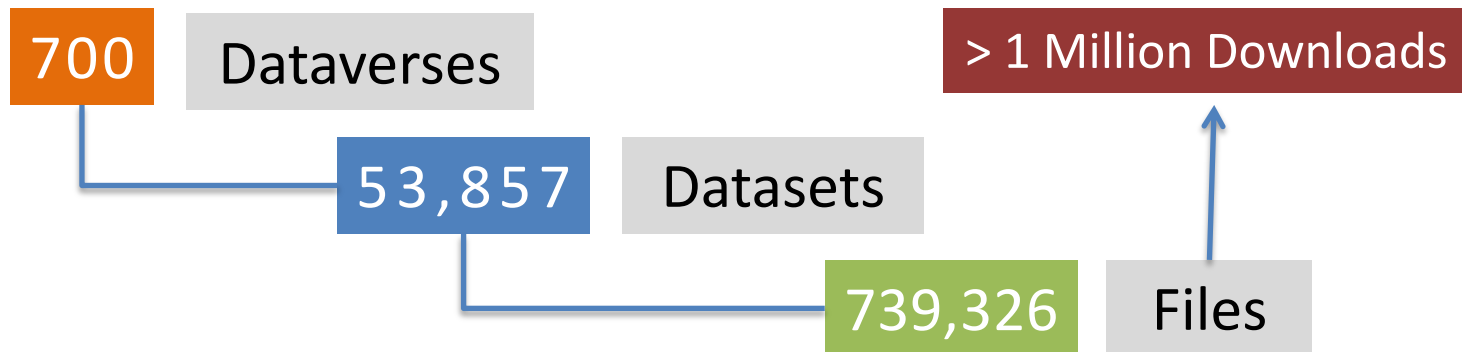
Software framework for publishing, citing and preserving research data (open source on [github](https://github.com) for others to install)

Provides incentives for researchers to share:

- Recognition & credit via data citations
- Control over data & branding
- Fulfill Data Management Plan requirements

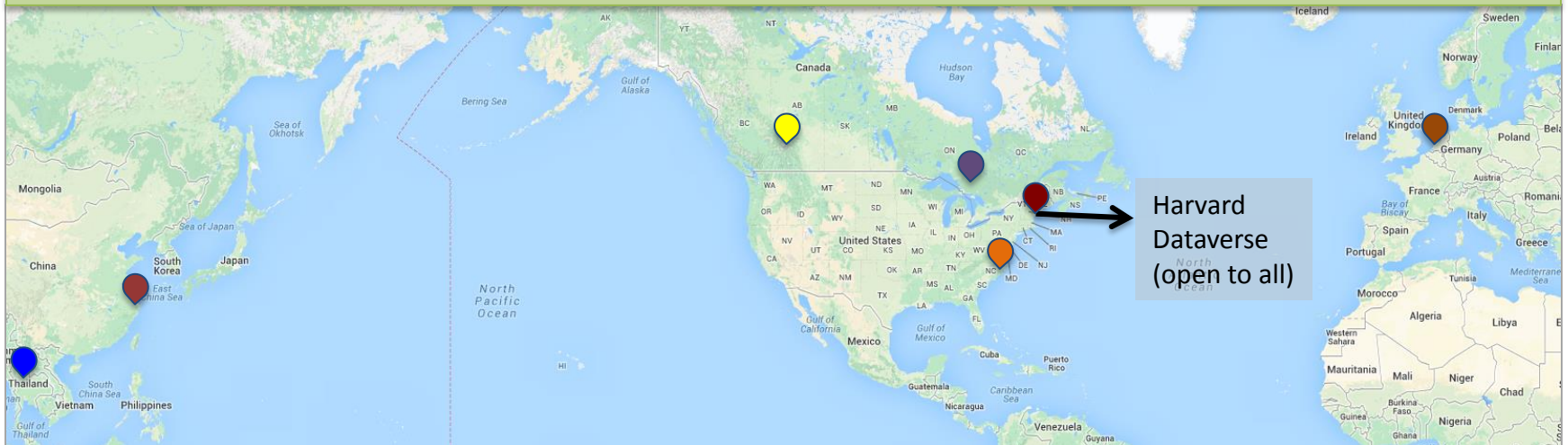


Harvard Dataverse (open to all repository instance at Harvard) currently has:



Who's Using Dataverse?

Worldwide Dataverse Installations



Types of Dataverses (across all research domains)

Institutions

(ODUM, MIT, OCUL,...)

Journals

(AJPS, Open Health Data,...)

Projects

(IFPRI, PSI, COMPLETE,...)

Researchers

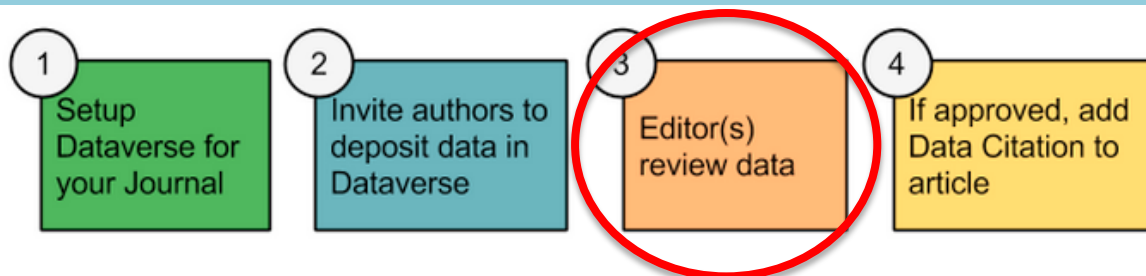
(Jonathan McDowell, Eric Dunipace,...)

Journals Working With Dataverse

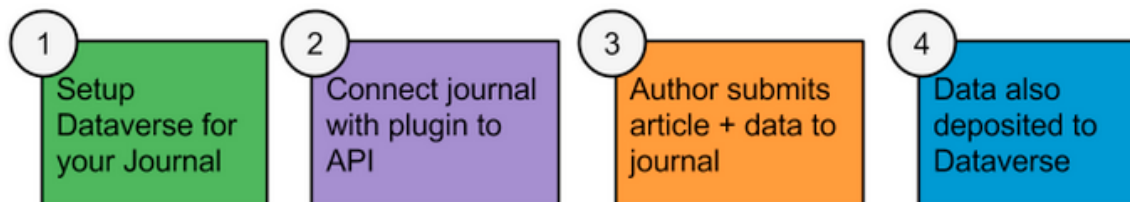
Option A. Journals include Dataverse as a Recommended Repository



Option B. Authors Contribute Directly to a Journal Dataverse



Option C. Seamless Integration btw Journal + Dataverse (e.g., OJS)



OJS Journal



Citation
to Data



Citation
to Article

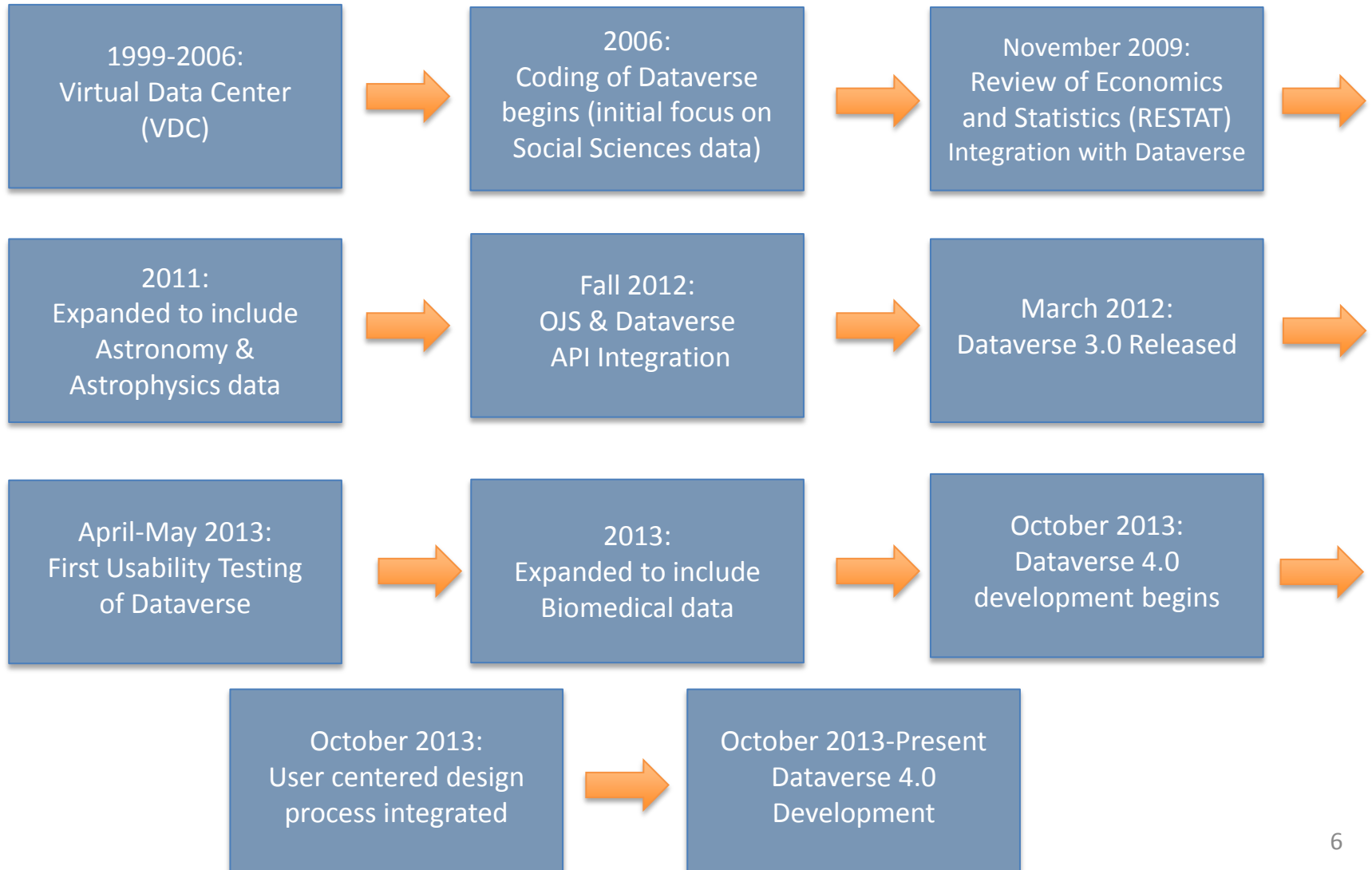
Journal Dataverse



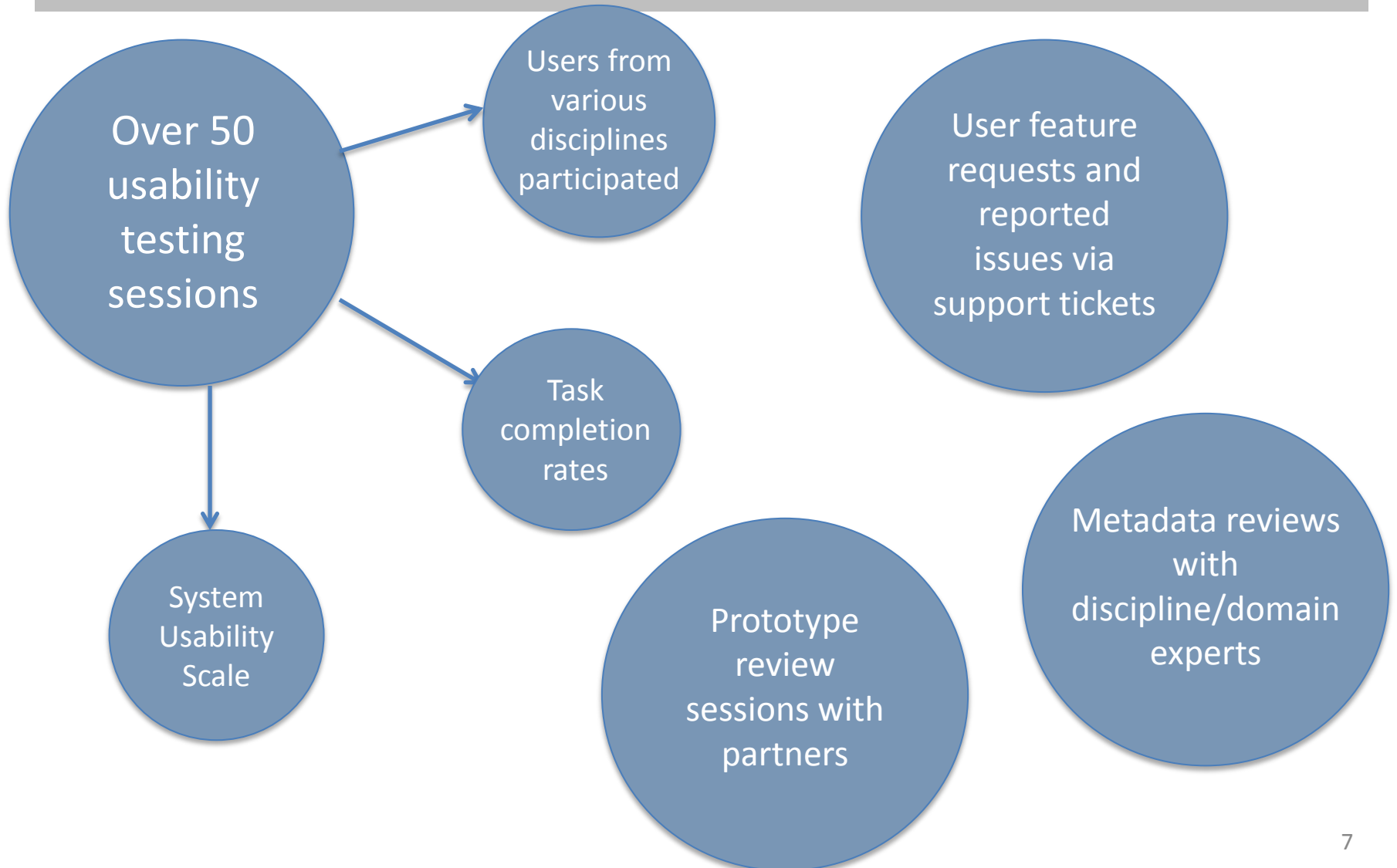
Details/Updates:


- Integrating w/ PKP's Open Journal Systems (Data Deposit API).
- Pilot with ~ 50 journals + expanding outreach.
- OJS Dataverse plugin now available with latest OJS release.
- **Future:** Embed Dataverse widget into journal article.


Dataverse Milestones



How our users have influenced 4.0:



 **Dataverse** Beta
[Q](#) [About](#) [Support](#) [Sign Up](#) [Log In](#)



Harvard Dataverse

Harvard Dataverse [✉ Email Dataverse Contact](#)

The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

[Q Find](#) [Advanced Search](#)
[+ Add Data](#)

[Dataverses \(21\)](#)

[Datasets \(24\)](#)

[Files \(7\)](#)

1 to 10 of 45 results

Sort ▾

< Previous 1 2 3 4 5 Next >

Overview of Dataverse 4.0

Affiliation

Harvard University (13)

COMPLETE (3)

California Institute of Technology (3)

University of Colorado (3)

University of Texas (3)

[More...](#)

Publication Date

2014 (45)

Author Name

King, Gary (6)

COMPLETE team (3)

Enoch, Melissa L. (3)


Evans II, Neal J. (3)

Glenn, Jason (3)

Harris Interactive, Inc., 2014, "TEST Harris 2008 Public Opinion Survey, study no. 35884", <http://dx.doi.org/10.5072/FK2/174>, Harvard Dataverse, V1

Topics addressed in this study include the digital TV conversion, confidence in the leaders of major institutions, health behaviors (specifically weight, seat belt use, and smoking habits), attitudes...


[Odum Institute Test Dataverse](#) (Odum Institute)

 May 23, 2014

This is the test DVN for Odum Institute.

[Preview Recently Released Datasets \[-\]](#)

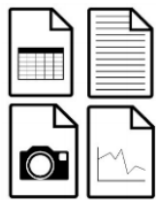
[Peking University Library Research Data Management Dataverse](#) (Peking University Library)

 May 21, 2014

the research data management group of peking university library

Try our Beta site: <http://dataverse-demo.iq.harvard.edu/>

Rigorous Data Publishing Workflows



Upload

Draft Dataset

Note: A Published Dataset **cannot** be deleted (only deaccessioned, if legally needed).

Publish Version 1

Authors, Title, Year, DOI, Repository, **V1**

Published Dataset v1

Publish Version 1.1: small metadata change; citation doesn't change.

Published Dataset v1.1

Publish Version 2: File change (automatic); big metadata change; or citation changes.

Authors, Title, Year, DOI, Repository, **UNF, V2**

Published Dataset v2

Expanding Metadata Support

Metadata Schema	Version 3.6	Version 4.0
DDI (General & Social Science)*	X (v2.1)	X (v.2.5)
Simple Dublin Core	X	X
Dublin Core Terms		X
DataCite 3.0		X
Virtual Observatory (Astrophysics)**		X
ISA-Tab (Biomedical)***		X

* Including variable level metadata found in [tabular data files](#).

** Automatically extracts relevant metadata from the header [FITS files](#).

*** Controlled vocabulary maps to ontologies/taxonomies (OBI, NCBI,...).

Design Type

- Case Control
- Cross Sectional
- Not Specified
- Parallel Group Design
- Perturbation Design

Biomedical Metadata

Factor Type

- Age
- Biomarkers
- Developmental Stage
- Cell Surface Markers
- Cell Type/Cell Line

Measurement Type

- DNA Methylation Profiling (Bisulfite-Seq)
- DNA Methylation Profiling (MeDIP-Seq)
- Histone Modification (ChIP-Seq)
- Protein-RNA Binding (RIP-Seq)
- Transcription Factor Binding (ChIP-Seq)

Organism

- Danio rerio
- Homo sapiens
- Mus musculus
- Rattus norvegicus

Cell Type



Enhanced Faceted Search



Email Dataverse Contact

The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

Search this Dataverse...

Find

Advanced Search

Add Data

Datasets (21)

Files (57)

Datasets (24)

Affiliation

Harvard University (13)

COMPLETE (3)

California Institute of

Technology (3)

University of Colorado (3)

University of Texas (3)

More...

Publication Date

2014 (45)

Author Name

King, Gary (6)

1 to 10 of 45 results

Sort

< Previous

1

2

3

4

5

Next >

TEST Harris 2008 Public Opinion Survey, study no. 35884



May 23, 2014 Odum Institute Test Dataverse

Harris Interactive, Inc., 2014, "TEST Harris 2008 Public Opinion Survey, study no. 35884", <http://dx.doi.org/10.5072/FK2/174>, Harvard Dataverse, V1

Topics addressed in this study include the digital TV conversion, confidence in the leaders of major institutions, health behaviors (specifically weight, seat belt use, and smoking habits), attitudes...

Odum Institute Test Dataverse (Odum Institute)



May 23, 2014

This is the test DVN for Odum Institute.

[Preview Recently Released Datasets \[+\]](#)


Enhanced Faceted Search

The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

election

Find Advanced Search

+ Add Data

-  **Dataverses (1)**
-  **Datasets (3)**
-  **Files (0)**

Affiliation

Harvard University (2)
IQSS (1)

Publication Date

2014 (4)

Author Name

King, Gary (2)
Adams, Greg (1)
Altman, Micah (1)
Benoit, Kenneth (1)
Gay, Claudine (1)

More...

Author Affiliation

Harvard University (2)


Keyword

congressional districts (1)
election (1)
election districts (1)
election returns (1)

1 to 4 of 4 results


Sort

Election Data Dataverse (IQSS)

 May 16, 2014
Contains various **election** datasets.


Preview Recently Released Datasets [+]

Elections to the United States House of Representatives, 1898-1992

 May 8, 2014 Gary King Dataverse
King, Gary, 2014, "Elections to the United States House of Representatives, 1898-1992", <http://dx.doi.org/10.5072/FK2/84>, Harvard Dataverse, V1

... and the effect of party incumbency on **election** outcomes, contains **election** returns for **elections** to the United ...

Election Data from 1850-1923 in Georgia

 May 16, 2014 Election Data Dataverse
Smith, Jon, 2014, "Election Data from 1850-1923 in Georgia", <http://dx.doi.org/10.5072/FK2/139>, Harvard Dataverse, V1

... Data spanning from 1850-1912 for the state of Georgia. Includes state and local **election** data. ...

Keyword: **election**

Record of American Democracy, All Key Data Files

 May 8, 2014 Gary King Dataverse

Enhanced Faceted Search

The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

election

Find Advanced Search

Add Data

- Dataverses (0)
- Datasets (1)
- Files (0)

Author Name: King, Gary Subject: Social Sciences Author Name: Altman, Micah Keyword: election returns

1 to 1 of 1 result

Sort

Publication Date
2014 (1)

Author Name
Adams, Greg (1)
Altman, Micah (1) ✕
Benoit, Kenneth (1)
Gay, Claudine (1)
King, Gary (1) ✕

More...

Keyword
congressional districts (1)
election districts (1)
election returns (1) ✕

Subject
Social Sciences (1) ✕

Record of American Democracy, All Key Data Files

May 8, 2014 Gary King Dataverse

King, Gary; Palmquist, Bradley; Adams, Greg; Altman, Micah; Benoit, Kenneth; Gay, Claudine; Lewis, Jeffrey B.; Mayer, Russ; Reinhardt, Eric, 2014, "Record of American Democracy, All Key Data Files", <http://dx.doi.org/10.5072/FK2/71>, Harvard Dataverse, V1

... The Record of American Democracy (ROAD) data provide **election** returns, socioeconomic summaries ...
Keyword: **election** districts

Expanded Advanced Search

Dataverses

Name	<input type="text"/>
Affiliation	<input type="text"/>
Description	<input type="text"/>

Datasets: Citation Metadata

Datasets: Social Science and Humanities Metadata

Datasets: Astronomy and Astrophysics Metadata

Datasets: Biomedical Metadata

Files

Name	<input type="text"/>
Description	<input type="text"/>
File Type	<input type="text"/>
Variable Name	<input type="text"/>
Variable Label	<input type="text"/>

Find

Ability to search on specific dataset metadata fields across various domains



- Integrated with Dataverse & Zelig
- For users at all statistical levels
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets

Dataverse **Beta** Q About Support - Sign Up Log In

[Election Data Dataverse](#) (IQSS)

[Harvard Dataverse](#) > [Election Data Dataverse](#) > [Election Data from 1850-1923 in Georgia](#) ✉ Email Dataset Contact




Election Data from 1850-1923 in Georgia

Smith, Jon, 2014, "Election Data from 1850-1923 in Georgia", <http://dx.doi.org/10.5072/FK2/139>, Harvard Dataverse, V1

Data spanning from 1850-1912 for the state of Georgia. Includes state and local election data.

Keyword	election
Subject	Social Sciences

[Files](#) [Metadata](#) [Versions](#)

 fearonLaitinData.tab Tabular Data, UNF:6:FILEFILEFILEFILE	 Explore	 Download
---	--	---

© Copyright 1997-2014, President & Fellows Harvard University.



- Integrated with Dataverse & Zelig
- For users at all statistical levels
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets

The screenshot displays the TwoRavens web interface. On the left, a sidebar shows the variable 'ccode' with its summary statistics: Mean: 451, Median: 451, Mode: NaN, Stand.Dev: 248, Minimum: 2, Maximum: 95, Valid: 6610, and Invalid: 0. Below the summary is a histogram of the 'ccode' variable. The main panel shows a causal diagram with three nodes: 'country' (blue circle), 'ccode' (blue circle), and 'cname' (orange circle). Arrows point from 'country' to 'ccode' and from 'ccode' to 'cname'. The 'ccode' node is surrounded by a circular ring of labels: 'TYPE', 'COUNTRY', 'DRIVER', and 'COUNTRY'. On the right, a 'Results Table' is visible, showing a list of models: gamma, logit, ls, negbin, poisson, and probit.



- Integrated with Dataverse & Zelig
- For users at all statistical levels
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets

The screenshot displays the TwoRavens web interface. On the left, a panel titled 'fearonLaitin' shows a list of variables under the 'Variables' tab. The variable 'war' is highlighted in blue. In the center, a causal diagram shows a yellow circle labeled 'lpop' with a dashed arrow pointing to a blue circle labeled 'war'. On the right, a 'Results Table' panel shows two diagnostic plots: 'Expected Values: E(Y|X)' and 'Predicted Values: Y-X', both displaying normal distribution curves. The interface includes buttons for 'Estimate', 'Force', 'Reset', 'Models', 'Set Cover.', and 'Results'. At the bottom left, the text 'ls(war ~ lpop)' is visible.



What to Expect After 4.0

WorldMap Integration

1. Upload a file containing geographic data into Dataverse
2. Easily visualize the data on the WorldMap system
3. WorldMap layer embedded into dataset in Dataverse

Email Dataset Contact

Social Disorder and Crime in Boston

BARI, 2014, "Social Disorder and Crime in Boston", <http://dx.doi.org/10.5072/FK2/40>, Root Dataverse, V1

Displays the distribution of five measures of social disorder across census block groups (CBGs) in Boston, MA: Social Disorder, Interpersonal Violence, Guns, Alcohol, and Social Conflict. The measures are derived from calls received by Boston's 911 Emergency Alert System in 2012.

Keyword social disorder, crime

Subject Social Sciences

Files

Metadata

Versions



social_disorder_in_boston_yqh.zip

ZIP Archive, MD5: 1a24e5f46b0a0e3211cbb6a9a107eb8ab

Map It!

Download

The .zip file contains a geospatial shapefile

Dataverse recognizes the shapefile. After background processing, the user sees a button that connects to the WorldMap

WorldMap Integration

1. Upload a file containing geographic data into Dataverse
2. Easily visualize the data on the WorldMap system
3. WorldMap layer embedded into dataset in Dataverse

Description

Displays the distribution of five measures of social disorder across census block groups (CBGs) in Boston, MA: Social Disorder, Interpersonal Violence, Guns, Alcohol, and Social Conflict. The measures are derived from calls received by Boston's 911 Emergency Alert System in 2012.

Keyword

social disorder, crime

Subject

Social Sciences

Depositor

[Redacted]

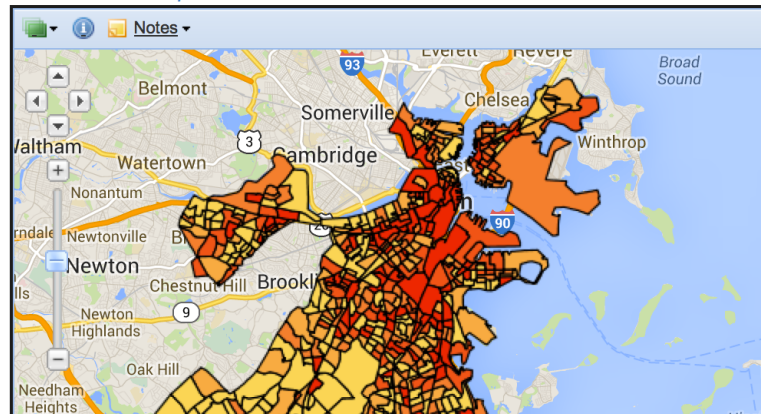
Deposit Date

2014-05-19

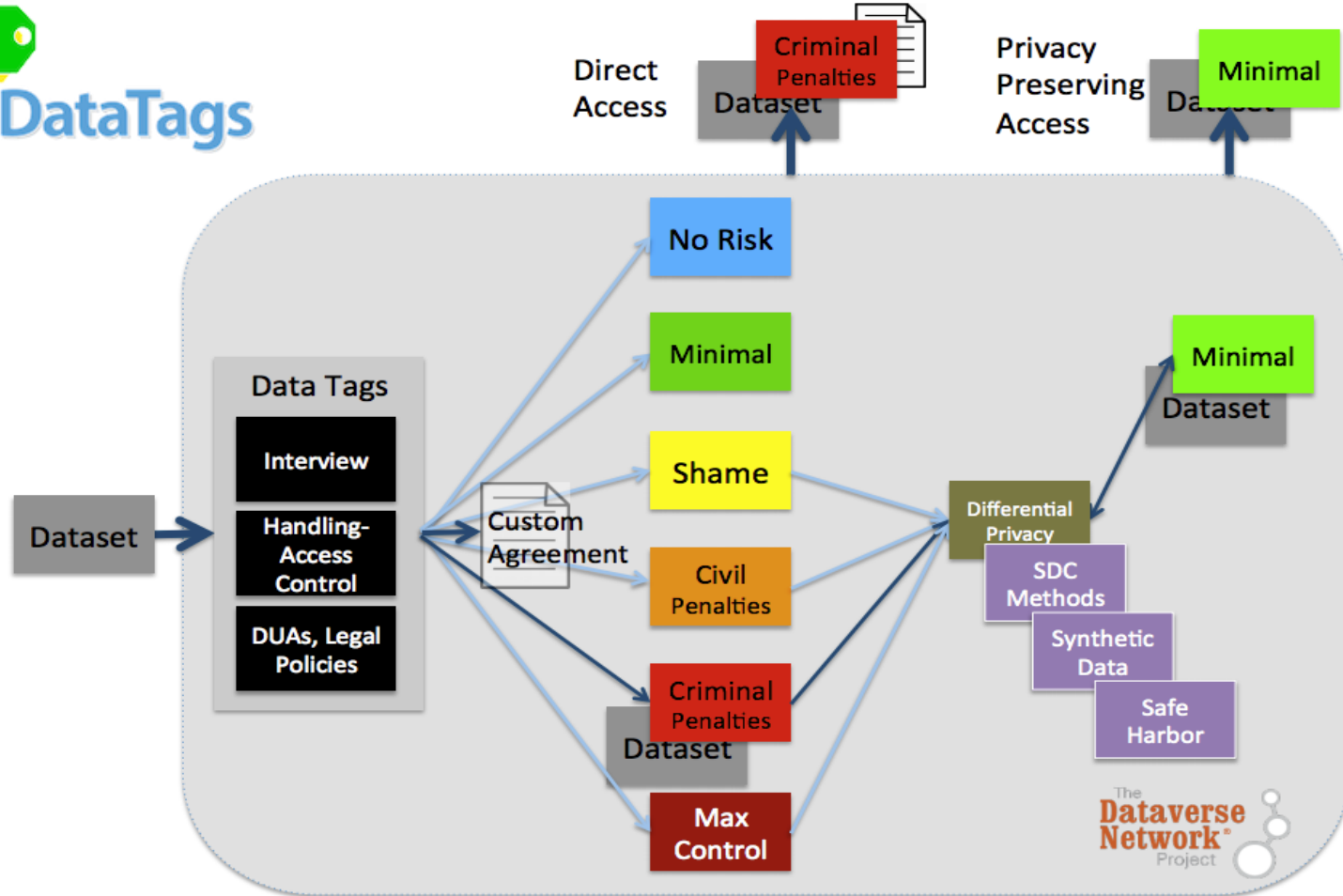
WorldMap Layer

[View on WorldMap](#)

**Show WorldMap link and
embedded layer on Dataverse**



DataTags: Sharing Sensitive Data



Demo: <http://datatags.org/>



Berkman
The Berkman Center for Internet & Society
at Harvard University



Data Tags

Sharing data with confidence

Start Tagging

Harm Levels, and Their Appropriate Tags

Level	D.U.A. Agreement Method	Authentication	Transit Encryption	Storage Encryption
NoRisk	None	None	Clear	Clear
Minimal	None	Email_or_OAuth	Clear	Clear
Shame	ClickThrough	Password	Encrypted	Encrypted
CivilPenalties	Sign	Password	Encrypted	Encrypted
CriminalPenalties	Sign	TwoFactor	Encrypted	Encrypted
MaxControl	Sign	TwoFactor	DoubleEncryption	DoubleEncryption

Final tags may not match the tags of a specific harm level. Hover over the terms to view an explanation.



Data Tags Sharing data with confidence

Person-specific

Does your data include personal information?

YES NO

Data Tags

DUA Agreement Method	n/a
Authentication Type	n/a
Transit Encryption Type	n/a
Storage Encryption Type	n/a

Full Interview

✔ Tagging Complete!

Direct Data Access

CriminalPenalties

DUA Agreement Method	Sign
Authentication Type	TwoFactor
Transit Encryption Type	Encrypted
Storage Encryption Type	Encrypted

Collaborations & More

Biomedical Metadata + Tools

- HSPH: TB Genomics & Molecular Epidemiology Data
- Harvard Stem Cell Institute
- FAIRport Data

Astronomy Metadata + Tools

- Seamless Astronomy Group
- Harvard-Smithsonian Center for Astrophysics

Also working on:

- Integration w/: ORCID, Open Science Framework, DataUp, DataBridge, ...
- Support for large-scale datasets with efficient data storage.

Data Science Team

Collaborations Resources Blog Team Search

Data Science
Research Frameworks for Data-Intensive Science,
Analytical Tools and Data Stewardship

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

Zelig Dataverse TwoRavens DataTags Consilience RBuild

About Us
Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation. Meet our team.

CURRENT EFFORTS
Reproducible and Reusable Science
Connecting research results to the underlying data and analysis is central to the validation and extensibility of scientific discoveries. Our tools encourage open data practices, and promote and enable data citation.

COMPUTATIONALLY ASSISTED EXPLORATION
We build analytical tools, such as Consilience and TwoRavens, that assist a

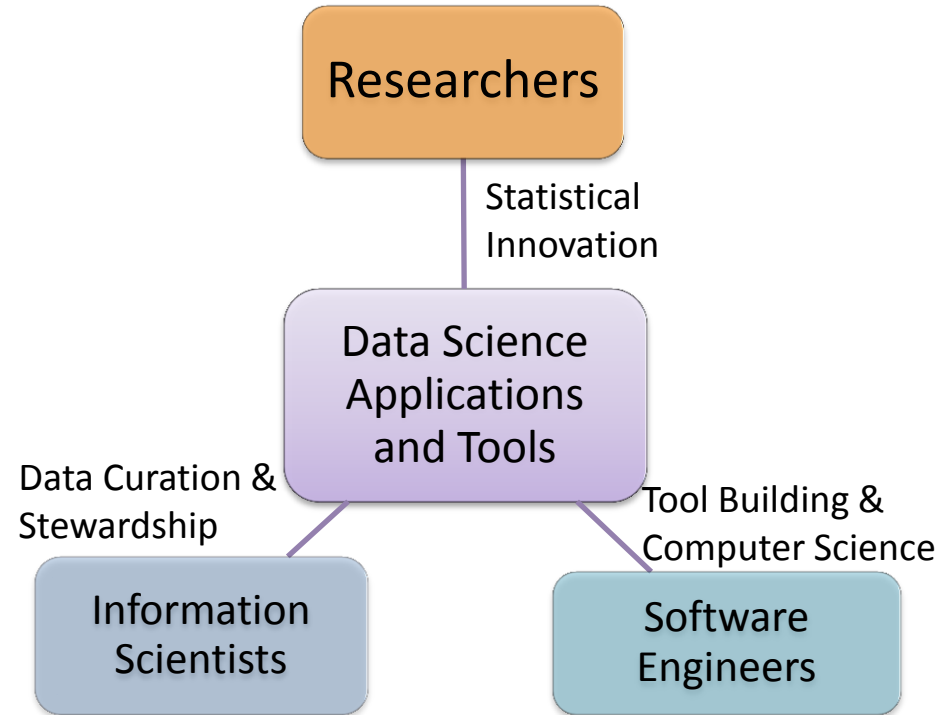
SOFTWARE PROJECTS
Zelig
Zelig is an interface, that allows a large body of different statistical models in the R statistical language to be implemented and interface.

THE DATVERSE NETWORK PROJECT

DATA SCIENCE BLOG
Dataverse 4.0 Beta
May 8, 2014

The Dataverse team has been hard at work on an extensive rewrite of the Dataverse application. Thanks to helpful feedback

<http://datascience.iq.harvard.edu>



Thank you!

Contact: ecastro@fas.harvard.edu; equigley@iq.harvard.edu